CECIL: A Database For Storing And Retrieving Clinical And Molecular Information On Patients With Alport Syndrome

Prakash M. Nadkarni[1*], Stephen T. Reeders[2,3] and Jing Zhou[2]
[1]Center for Medical Informatics, [2]Department of Human Genetics and [3]Howard Hughes Medical Institute, Yale School of Medicine, New Haven, CT 06510, USA

CECIL is a database that stores clinical and molecular information on patients with Alport syndrome. The clinical component of CECIL is specific to Alport syndrome; the component that stores and manipulates molecular data can be used for any disease caused by a gene mutation, such as cystic fibrosis. While offering the ability to retrieve patient data through compound Boolean queries, CECIL also offers the ability to manipulate sequence information in various ways. In particular, CECIL can perform an augmented sequence alignment of an abnormal (patient) DNA sequence with a reference sequence. CECIL is currently being used by members of the International Alport Syndrome consortium. We describe CECIL's features and discuss the design decisions made in generalizing CECIL's architecture.

INTRODUCTION

Alport syndrome [1] is a congenital disease manifesting as hereditary hematuric nephritis, with renal failure occurring anywhere from the first to the fifth decade of life. Associated findings may include sensorineural deafness, lenticonus, and abnormal platelet morphology. Most cases of Alport's are due to a spontaneous mutation in the gene for the alpha-5 chain of type IV collagen (HUMCOL4A5) [2,3,4], which is a component of basement membranes such as those in the glomerulus and the lens. This gene lies on chromosome X, and mutates at a frequency estimated at 1:10,000. Because of the X-linked nature of the mutation, the vast majority of affected individuals are males; females generally have a considerably more benign clinical presentation.

HUMCOL4A5 is assembled from 51 separate components (**exons**) [5], which are separated by non-coding stretches of DNA (**introns**). A mutation anywhere in the gene may cause the syndrome; the severity of the clinical manifestations depends on the degree of functional impairment in the mutated HUMCOL4A5 molecule. To date, more than 60 different Alport mutations have been recorded. These range from substitution of a single nucleotide to the absence of the entire gene [6,7].

The diagnosis of Alport syndrome is based on the clinical presentation and the family history. It may be confirmed by electron microscopy of biopsied renal tissue; the renal glomerular basement membrane shows a distinctive splitting and lamellation. The first step towards diagnosing the nature of an X-linked Alport mutation is digestion of the patient's DNA with a restriction enzyme. The enzyme breaks the DNA into smaller fragments that are electrophoretically separated. This followed by treatment (**hybridization**) of the digest with labeled DNA (a **probe**) that is known to span part of the HUMCOL4A5 region. Chromatograms (**Southern blots**) of patient DNA processed with different enzyme-probe combinations are then compared visually with corresponding chromatograms of normal DNA to look for the absence of normal fragments, or the appearance of abnormal fragments. If the Southern blot is not diagnostic, as happens sometimes when there is substitution of only a single nucleotide, all 51 exons in the patient may be amplified and compared against all the corresponding reference exons. This procedure, currently limited to research purposes only, is cheaper than renal biopsy plus electron microscopy; forthcoming advances in robotics may lower its cost dramatically.

The International Alport Syndrome Consortium, a voluntary association of researchers, accumulates clinical and molecular data on Alport syndrome and facilitates early information interchange through a newsletter. (The second and third authors are members of the consortium and the co-organizers of the Second International Conference on Alport Syndrome, held in New Haven in February 1993.) CECIL has been built to store data submitted to the consortium by individual investigators, and assist in its analysis. In particular, CECIL has been designed for correlation of clinical features (**phenotype**) with the molecular nature of the mutation (**genotype**) as data accumulates; this will be useful in providing genetic counseling. As a side effect, CECIL facilitates electronic newsletter publication. CECIL is named after Dr. Cecil

Alport, who first described the clinical syndrome. It runs on the Apple Macintosh and is implemented with the database package 4th Dimension® (4D), marketed by ACI.

DATA STORED IN CECIL

Information currently stored in CECIL falls into the following categories:

- clinical and clinical laboratory findings in affected patients ;
- pedigree information, to record the nature of the disease in individuals related to the patient (these patients may or may not have been entered into the database);
- DNA sequences obtained from patients;
- reference DNA sequences (the individual exons of HUMCOL4A5 as well as the complete cDNA sequence of HUMCOL4A5;
- mutations of these DNA sequences as recorded in patients;
- Southern blot information for both patient and normal DNA;
- citations of published work in Alport syndrome, and abstracts of that work;
- Rolodex®-type information on members of the Alport Syndrome consortium and other investigators.

The laboratory information that can be stored in CECIL includes images such as light and electron microscopy. Digitized images of both reference and patient Southern blots may be stored to facilitate visual comparison of the two if necessary.

FEATURES OF CECIL

Clinical Data

The clinical component of CECIL is fairly standard, allowing entry and editing of patient data as well as retrieval on compound Boolean criteria. Each patient is tagged with the name of the reporting investigator to encourage submissions and protect intellectual property. Patients are identified by coded identifiers to protect confidentiality. The choice of fields has been made by the second and third authors (who are domain experts) and refined by beta-testing within the Alport consortium community.

DNA Sequence Data

All DNA sequences entered into CECIL can be manipulated in limited ways. The manipulations include pretty-formatting , reverse-complementing, translation into a desired reading frame, and search for regular expressions The last uses an algorithm described by Kernighan and Plauger [7].

CECIL can align a sequence obtained from a patient with a reference sequence. The alignment program is a modification (by the first author) of William R. Pearson's ALIGN program (part of the widely used FASTA package for sequence manipulation [8]. The modified alignment will not only show the alignment of the two nucleotide sequences, but will also show the consequence of the mutation, in terms of the effect on the corresponding amino acid sequence.

Due to the ability of the cell to perform splicing of abnormal DNA, as well as post-translational modification of amino-acid sequences, the translation information does not always predict the mechanism of mutation correctly, but it does facilitate interpretation of the mutation in most cases.

When a mutated sequence is entered into the database, rather than enter the sequence itself in full, the user enters a series of **features**, which are the differences between the mutated sequence and the reference sequence. The number of features rarely exceeds three; most mutated sequences have but a single feature.

CECIL can compose a text description of a mutated sequence based on information that has been entered by the user. If possible, it will also reconstruct the mutated sequence from the reference sequence and a list of features. Such reconstruction is not possible in rare circumstances where it is known that a large segment of DNA has been deleted from the gene product, but the exact span of the deleted part (and its sequence) has not been worked out at the time of entry of the mutation into the database. The algorithm used for reconstruction of the mutated sequence is similar in principle to the UNIX utility *diff*, which generates a list of differences between two text files [9].
The user can also call up all mutations of a given type, and all patients with mutations showing a particular pattern, so that the clinical features for a particular class of mutations can be looked at.

Southern Blot Data

When looking at a patient's Southern blot data, CECIL will display the normal pattern for the particular probe-enzyme combination, if this has been entered previously. The list of fragments in the reference gel will be displayed next to the list of fragments in the patient, so that differences may be readily visualized.

Pedigree Information

CECIL is not intended to be a full-fledged pedigree management program, as the pedigree information stored within it is archival. Published papers reporting cases of Alport syndrome typically represent the pedigree as a drawing, and the user can enter a publication-quality drawing directly into CECIL.

The drawing of the pedigree is managed by a 4D add-on called 4D DRAW® (MicroCad Corporation), which lets the user draw a picture with standard tools in a user-friendly drawing environment. The drawing can be annotated with text (such as the coded identifiers of patients). Since several patients from the same pedigree can be stored in the same database, CECIL offers a point-and-click method of looking at patient data from the pedigree drawing, as illustrated in Figure 2.

Submitting data to CECIL

The individual investigator using CECIL to enter his or her own data can submit this data to the consortium chairman electronically through a submission facility. Since most members of the Alport consortium do not have access to the Internet, CECIL's submission tool currently creates a data file which can be copied to a floppy. This data can then be used to update the central copy of CECIL, which will be mailed to users at periodic intervals. We also intend to make the CECIL program and data available through anonymous *ftp* (file transfer protocol).

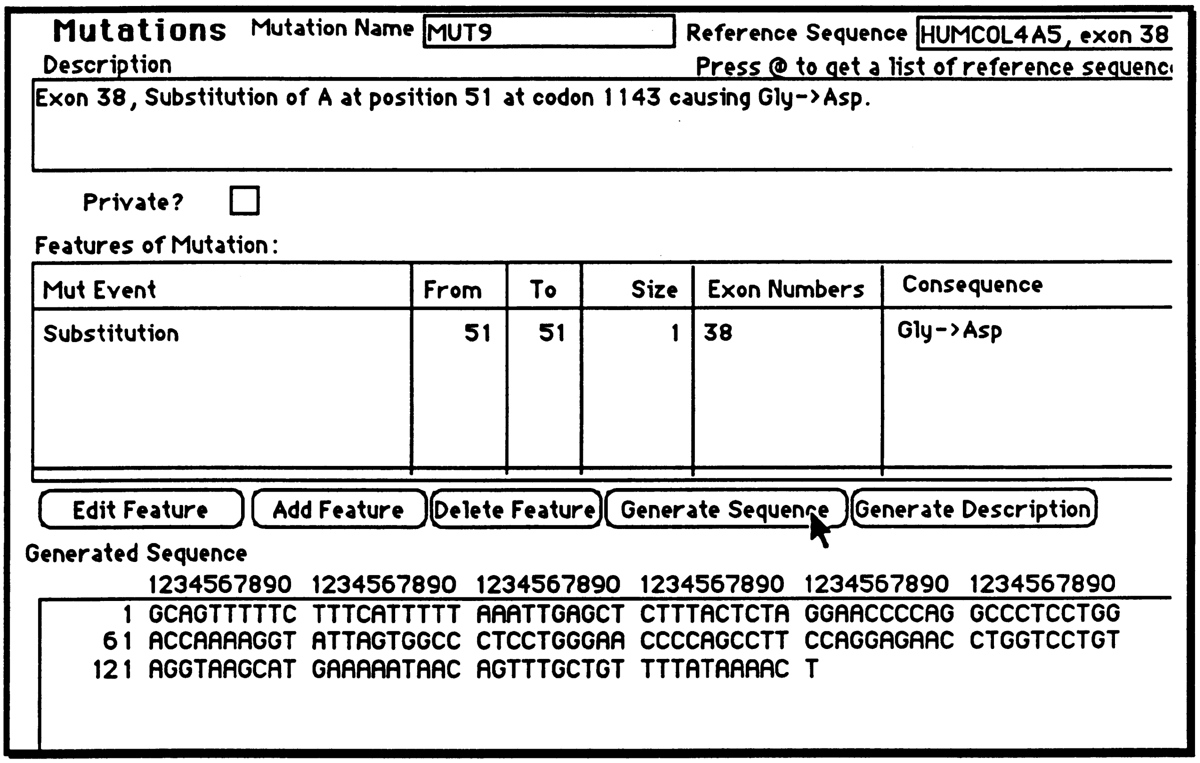

Figure 1: The user has entered the features of a mutation, defining the nucleotide involved (51) and exon(38). By clicking on "Generate Sequence" the mutant sequence is created. Clicking "Generate Description" produces a description of the mutation using a uniform syntax.

GENERALIZING CECIL: DESIGN DECISIONS

CECIL has been built in a fashion that allows straightforward porting to other mutation disorders such as cystic fibrosis. The molecular biology and the pedigree components of CECIL are directly portable. The only component requiring redesign for a different disease is that which is disease-specific; namely, the clinical and clinical laboratory component.

We have deliberately chosen to make the clinical component of CECIL non-portable. Such data is most efficiently stored as a series of fields corresponding to findings that are directly relevant to the disease. These fields may be two or three-valued Booleans (e.g., Yes/No/Not Available), enumerated (one of a limited set of choices), numeric, or string.

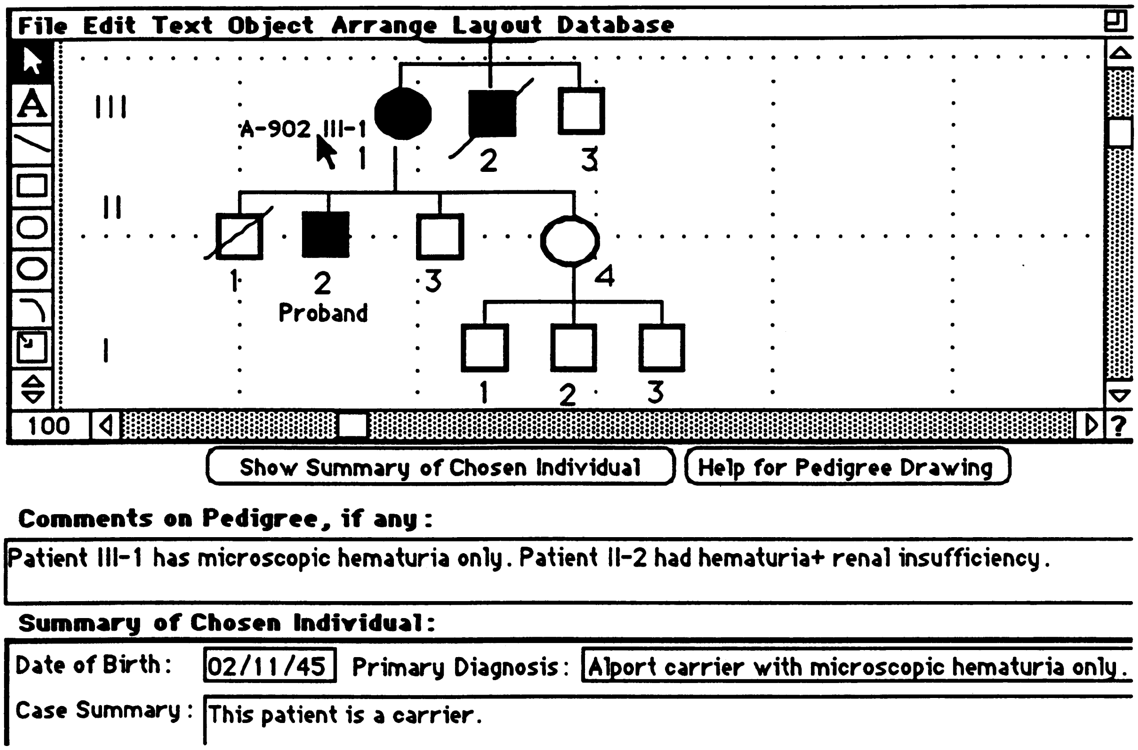

Figure 2: CECIL's Pedigree Display and Drawing Tool. The user has clicked on the identifier of an individual (A-902,III-1), and then clicked on "Show Summary of Chosen Individual". CECIL does an indexed search of the database on this identifier, showing a brief summary of the patient in a window.

For example, we store lenticonus as a Boolean field, hematuria (absent, microscopic or gross) as an enumerated field, and the 24-hour urine protein (in milligrams) as a number. (Some data is also pictorial, e.g., electron micrographs).

Storing the significant disease-specific attributes of the syndrome as individual fields rather than as LISP-style association pairs (finding-value) it is possible to do very efficient (indexed) compound Boolean searches. This functionality, an essential requirement of CECIL, would not be possible if such data were stored as free text.

While it is possible to build a general-purpose "data dictionary" facility that would let a non-computer-sophisticated individual describe the clinical component of a disease by describing the fields and their attributes, user-friendly packages such as 4th Dimension let the user perform such a process quite easily. The process is simplified because of the archival nature of the clinical information in a mutations database; it can almost always be stored in a single database table.

Secondly, the esthetics of the screens used for eliciting and displaying clinical information are best left to the individual designer. Screen design with graphically oriented databases such as 4th Dimension is intuitive and rapid. We were able to design the clinical component of CECIL in less than a day, and we expect that the experience of researchers working with other mutation disorders will not be different.

To let individual laboratories modify the clinical component of CECIL for a different disease, we distribute CECIL in source form.

PRESENT STATUS OF WORK AND FUTURE DIRECTIONS

Our work has been well accepted by the Alport consortium. CECIL currently holds 45 reference sequence sequences and 76 patients with unique mutations. Because of the rarity of Alport syndrome and the slow accumulation of interesting new cases, database performance is , and is likely to remain, a non-issue. However, for diseases like Cystic Fibrosis, which have many more mutations and patients, performance may require tuning as the database size increases.

Like all modern database engines, 4D allows such tuning through the specification of secondary indexes through a point and click interface. Such secondary indexing is particularly important in the performance of compound Boolean searches.

Our current implementation of CECIL lacks wide-area connectivity (4D lacks this functionality.) For large research consortia, we envisage a need for consortium members to connect to a (remote) central database via the Internet. Database engines such as Sybase possess such connectivity. 4D (as well as MS-Windows databases such as Microsoft Access) can act as front-ends ("clients") to a Sybase server. Such an application has the "look-and-feel" of the native platform (e.g., Windows or Mac), while the user is really operating on data residing on a different machine and operating system.

We have created a programmer's toolkit for building 4D clients to Sybase, called SQLGEN [11],which is in production use at Kenneth Kidd's genetics lab at Yale. We intend to employ SQLGEN in building a centrally accessible Sybase database for the consortium for Tourette's syndrome, of which Dr. Kidd is a member.

CONCLUSIONS

As research in diseases caused by mutations progresses, databases to manage clinical and laboratory data pooled from multiple sources are likely to become increasingly important. In storing data that can be analyzed in various ways, they can help to provide pointers the underlying mechanism of the disease and suggest the direction of future laboratory research.

CECIL straddles the gap between clinical and molecular biology databases by offering the functionality required of both. We expect that as patient and sequence data accumulates, the vital questions of phenotype-genotype correlation in Alport syndrome can be investigated and answered with CECIL.

Acknowledgments: During the course of this work, the first author was supported by Grant R01 HG00175 of the National Center for Human Genome Research.

References

[1]. Reeders ST. Molecular Genetics of hereditary nephritis. Kidney International 42:783-792, 1992

[2]. Hostikka SL, Eddy RL, Byers MG, Hoyhtta M, Shows TB & Tryggvason K. Identification of a distinct type IV collagen alpha chain with restricted kidney distribution and assignment of its gene to the locus of X-linked Alport syndrome. Proc. Natl. Acad. Sci. USA 87:1606-1610, 1990.

[3] Zhou J, Herz, JM, Leinonen A, Tryggvason K. Complete amino acid sequence of the human α-5(IV) collagen chain and identification of a single base mutation in exon 23 converting glycine-521 in the collagenous domain to cysteine in an Alport syndrome patient. J Biol Chem 267,12475-12481, 1992.

[4] Zhou J, Hostikka SL, Chow LT and Tryggvason K. Characterization of the 3' half of the human type IV collagen α-5 gene which is affected in Alport syndrome. Genomics 9, 1-9, 1991.

[5]. Zhou J, Hostikka SL, Chow LT and Tryggvason K. Structure of the Human COL4A5 gene for the Type IV collagen α-5 chain. (submitted)

[6] Zhou J, Barker D, Hostikka SL, Gregory M, Atkin C ,Tryggvason K. Single base mutation in α5 (IV) collagen chain converting a conserved cysteine to serine in Alport syndrome. Genomics 9, 10-18, 1991.

[7] Knebelman B, Druout L, Forestier L, Deschenes G, Grunfeld JP, Gubler MC, Antignac C. Mutations in the COL4A5 gene in families studied at Necker Hospital. Presented at the Second International Conference on Alport Syndrome, New Haven, CT, 1993.

[8]. Kernighan BW, Plauger PJ. Software tools in Pascal. Addison-Wesley, Reading, MA, 1981.

[9]. Pearson W.R. and Lipman D. Improved Tools for Biological Sequence Comparison. Proc. Natl. Acad.Sci USA **85**: 24444-2448, 1988.

[10]. UNIX User's Reference Manual. Department of Electrical Engineering and Computer Science, University of California, Berkeley. April 1986.

[11] Nadkarni, Prakash. SQLGEN Version 1.0 Manual. Yale Center for Medical Informatics, Yale School of Medicine, New Haven, CT. 1993.